DOCUMENT RESUME

ED 291 798 TM 011 101

AUTHOR Douzenis, Cordelia; Rakow, Ernest A.
TITLE Outliers: A Potential Data Problem.
PUB DATE Nov 87
NOTE 22p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Mobile, AL, November 11-13, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Data Analysis; *Data Interpretation; *Statistical Analysis; Statistical Distributions
IDENTIFIERS *Outliers

ABSTRACT

Outliers, extreme data values relative to others in a sample, may distort statistics that assume internal levels of measurement and normal distribution. The outlier may be a valid value or an error. Several procedures are available for identifying outliers, and each may be applied to errors of prediction from the regression lines for utility in a bivariate situation. Because the existence of outliers in a data set may bias the results, some researchers have proposed the use of robust statistics in the presence of outliers with trimming, bimean, or winsorization techniques. Six examples of the effect of outliers on statistical results are given with tables illustrating distortion in each situation. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED291798

OUTLIERS: A POTENTIAL DATA PROBLEM

by

Cordelia Douzenis

and

Ernest A. Rakow

Dept. of Foundations of Education
Memphis State University
Memphis, Tennessee 38152
Phone (901) 454-3420

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
CORDELIA DOUZENIS &
ERNEST A. RAKOW
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the Mid-South Educational Research Association in Mobile, Alabama, November 11-13, 1987.

TM011101

BEST COPY AVAILABLE

OUTLIERS: A POTENTIAL DATA PROBLEM

Cordelia Douzenis and Ernest A. Rakow
Memphis State University

Data analyzed in educational research and evaluation studies sometimes has extreme values or outliers. Outliers may distort statistics which assume interval levels of measurement and/or normal distribution. Statistics which include outliers when being calculated lead to inaccurate conclusions. This paper defines outliers, discusses procedures for identifying them, discusses statistical procedures for analyzing data in their presence, and provides some examples.

An outlier is an extreme data value relative to others in the sample. The outlier may be an extremely large or small value in a univariate distribution, or in a bivariate distribution, may be a value extremely deviant from the regression line. Such an outlier may be a valid value or an error.

Procedures for identifying univariate outliers may utilize quartiles and interquartile range (Tukey, 1977), the mean plus or minus a specified number of standard deviations (Sachs, 1984), or the ratio of sums of square deviations about the mean excluding outliers relative to the sum of squares including outliers (Grubs, 1950). Each of the above procedures may be applied to the errors of prediction from the regression line for utility in a bivariate situation. Barnett and Lewis (1978) indicate subjective judgment by a researcher is required when deciding whether or not an observation is an outlier.

The existence of outliers in a data set may bias the results. Bancroft and Han (1981), Hawkins (1980), and Tukey (1977) propose the use of robust statistics in the presence of outliers. For example, for the arithmetic mean a researcher might substitute the median, trimean, Winsorized mean, or trimmed mean.

The examples presented demonstrate provisions for outliers (1) in a component score when combining scores to form a composite, (2) when means are being calculated for use in time series analyses, (3) when sample values are being used to test hypotheses about means, and (4) when dealing with two variables jointly.

OUTLIERS: A POTENTIAL DATA PROBLEM

by

Cordelia Douzenis and Ernest A. Rakow

Overview

Data analyzed in educational research and evaluation studies sometimes contains extreme values or outliers. Outliers may distort statistics which assume interval level of measurement and/or normal distribution. Also, the presence of outliers may be an indicator of a weakness in the statistical model. Therefore, interpretation of statistical results for data that include outliers may lead to inaccurate conclusions.

This paper defines outliers, discusses procedures for identifying them, discusses statistical procedures for analyzing data in their presence, and provides some examples of the effects outliers can have on various types of data situations. The examples presented demonstrate provision for outliers 1) in a component score when combining scores to form a composite, 2) when means are being calculated for use in time series analyses, 3) when sample values are being used to test hypotheses about means, and 4) when dealing with two variables jointly.

Definition of Outliers

An outlier is an extreme data value relative to others in the sample. An outlier is "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett and Lewis, 1978). The outlier may be an extremely large or small value in a univariate distribution, or in a bivariate distribution may be a value extremely deviant from the regression line. Outliers or "wild shots" (Winer, 1971) are highly suspect observations in a sample indicating data points which are not typical of the rest of the sample. Such an outlier may be a valid value or an error.

Hawkins (1980) lists three situations in which an outlier could be considered a valid value:

1. It is an accurate value randomly selected from a normal distribution which happens to be an observation with very low probability of occurrence;
2. It is an accurate value selected from a skewed distribution or a heavily tailed distribution; or
3. It is an accurate value from a different population distribution than nearly all other values.

Errors may result from measurement or computational errors such that the sample reflects a population other than the one of interest. An invalid outlier could arise from such errors as data entry errors, coding errors, erratic errors, measurement errors, judgment errors, execution faults or computation errors (Sachs, 1984). When possible, the source of the error should be located and the outlier replaced with a correct value; otherwise the erroneous value should be deleted (Sachs, 1984). According to Lyon (1970), it is possible for large errors to occur occasionally due to the characteristics of the normal distribution. What one must determine is "what is the probability that so far out a value could occur by chance on the hypothesis that all values belong to the same normal population" (Lyon, p. 193).

Observations that depart considerably from the center of a distribution demand special consideration. Even if an extreme value can be considered valid, one must determine how to handle its presence in a data set. Any statistical tests or estimates of population parameters based on a sample containing such values must be questioned. Killeen (1985) states, "Few experimentalists would retain a data point that is 5 or 6 standard deviations from the mean, but what about one that is 3 standard deviations from the mean?" (p. 526).

Identification of Outliers

Many researchers consider extreme any value that exists outside the typical range of values for a particular variate. Sachs (1984) discusses a general rule that allows a value to be removed from the data set provided it lies at least four standard deviations from the mean after new statistics have been calculated without the inclusion of the suspect value. This "4 sigma region" includes 99.99% of the values for a normal distribution. He also suggests that statistical analyses might need to be carried out both with and without the extreme values. If the presence of the outlier affects the results, one should be cautious with interpretations.

Barnett and Lewis (1978) state that "rejection of an outlier turns a random sample into a censored one" (p. 32). Rey (1983) suggests that "rejecting an outlier drastically modifies the probability levels of all inferences" (p. 196). Many statisticians feel that with a large data base outlier tests should be standard. Although a statistical test can never give a definitive answer, one can test if the outlier was drawn from the same population the other data originated from.

Sachs (1984) states several reasons for using outlier tests including 1) routine inspection of the reliability of the data, 2) advisement of need to better control data collection, and 3) recognition of extreme data. The type of test needed depends on whether previous knowledge about the population which the data was drawn from is available. Lyon (1970) believes that a lack of information could have an effect on whether or not an extreme value is found to be significantly different. Sachs also mentions that removal of a significantly different value is contingent upon whether the data is approximately normally distributed.

Naus (1975) presents several statistical methods for detecting outliers including Nordbotten's gross error control method and Tukey's full normal

rejection (FUNOR) procedure. The gross error control method studies the internal consistency of variability of normally distributed data to check for outliers. It can be used when the mean and standard deviation of the sample are not known. A drawback to this method is that observations have to be very extreme to be considered outliers. Tukey's FUNOR procedure uses the median as a basis for detecting outliers. The FUNOR procedure only works when the assumption of normality holds.

In Exploratory Data Analysis Tukey (1977) proposed measures for identification of extreme values utilizing the interquartile range ("H-spread") and quartiles. Values 1.5 times the interquartile range (a step) below the twenty-fifth or above the seventy-fifth percentiles were considered "outside." (In a normal distribution these would be outside the range of Z values $\pm$ 2.70). True outliers or "far out" values are those beyond three times the interquartile range (two steps) below the twenty-fifth or above the seventy-fifth percentiles. (In a normal distribution "far out" values are outside the range of Z values $\pm$ 4.72.) In some situations these extreme values may provide hints for further investigation, while in others they maybe ignored when discussing the typical nature of that distribution (Tukey, 1977).

Grubbs (1950) developed a test for identifying outlying observations via comparison of two sums of squares. The first is the sum of squared deviations from the mean, excluding the suspected outliers from the mean and sum of squares. The second is, for all scores, calculate the sum of squared deviations from the mean of all scores. The ratio of the first divided by the second is compared to a special table derived by Grubbs to determine significance. This table is for samples of sizes three through twenty-five and for significance levels of .10, .05, .025, and .01. Low values for this ratio of sums of squares indicate the presence of outliers. Critical values of this ratio decrease with increasing sample size and also with decreasing probability values.

Robust Statistics

Robust statistical techniques have been developed to reduce the effect of outliers. Bancroft and Han (1981) define a robust estimator as "an estimator which performs well under a variety of underlying conditions" (p. 158). Barnett and Lewis (1978) define robust as "insensitivity to small deviations from the assumptions" (p. 45). Some of these techniques are trimming, bimean, and winsorization.

Trimming involves deleting an identical percentage of data from both the upper and lower tail. There are conflicting opinions as to how much of the data should be removed. Killeen (1985) states that a total of 5% should be trimmed from the data. Sachs (1984) suggests that a total of less than or equal to 3% should be removed. However, in cases of strong heterogeneity, Sachs says that one can trim up to 6% of the distribution.

Winsorization is the reassignment of extreme data values the value of their nearest neighbors. The winsorization procedure involves the value g which is the number of scores to be replaced. For example, if g = 3, then the three highest and the three lowest scores would be replaced by their next closest extreme. Winer (1971) states that the value of g should be tailored to the individual sample. One criticism of winsorization is that it makes partial use of the outliers. Barnett and Lewis (1978) also mention that both winsorization and trimming have distributional implications.

The bimean (Killeen, 1985) is a technique developed by Mosteller and Tukey that weights observations less as they depart from the middle of a distribution. This procedure utilizes the biweighted mean which is the weighted arithmetic mean of the data. The calculation of the bimean is iterative since it begins with an estimated value for the biweighted mean such as the median.

Another method of controlling outliers is Tukey's full normal modification

procedure (Naus, 1975). After one detects outliers with the full normal rejection procedure (FUNOR), the full normal modification procedure is then applied to the data. First, the values rejected by FUNOR are replaced with a median (or typical value). Second, the nonrejected but still extreme values are winsorized. Naus states that application of this procedure does not seriously distort the data.

The remainder of this paper presents six situations to demonstrate the effect of outliers upon statistical results.

Situation 1: Composite Score

In this situation, the goal was to calculate a composite score which was the mean of scores on several other variables. Outliers have the potential for causing a noticeable distortion in this composite score even after first converting all subscores to standard scores (Z). This application is actually modeled after results observed in analyzing data from the Tennessee Career Ladder. However, the data are purely fictitious.

For this situation, summary statistics were calculated for fifteen hypothetical variables. The first four variables were dichotomous, using values of 0 and 1. The next five variables ranged from 1 to 5. Two ranged from 1 to 9, while one ranged from 11 to 20 (but 0 was the minimum possible). The last three ranged from 1 to 100, but for the last one the lowest observed score was actually 80. For distributions with these ranges the means and standard deviations are shown in Table 1. Observe that when more than one variable has the same range, they are sequenced such that the means are in increasing order. Also observe, that when examining statistics for the same range, that standard deviations decrease as the mean shifts from the center of the distribution to a higher value, that is, as the distribution becomes negatively skewed. It may appear that some of these means and standard deviations are very extreme, but they are reasonably close to some values actually observed in the Tennessee Career Ladder data.

One approach used to combine scores with different variabilities into a composite has been to weight them equally by first converting all scores to standardized (Z) scores. Averaging such Z scores has been thought to allow each variable to be weighted equally. The Z score which would correspond to the minimum value in each distribution was calculated. In Table 1 it can be observed that the minimum Z's range from -1.00 to -15.75, indicating presence of outliers for some variables. The composite score is the mean on all fifteen variables. If a person scored at the mean on all fifteen variables, then his composite would be 0.00, i.e. the mean of fifteen Z's all equal to zero. If a person scored at the mean on fourteen of fifteen variables and had the minimum score on the other one, the values shown in the column labeled "Average: This Min Z, Others Mean" in Table 1 would be the composite Z corresponding to using that "Min Z." These composite scores range from -0.07 to -1.05, even though fourteen of the fifteen values averaged were zero. Apparently, the outlier for variable 15 made this range so large. This one variable is so extreme that a person with that one low score is well below all others. Also shown in Table 1 are the Z scores which would correspond to the maximum value for each distribution. These maximum Z's range from 0.06 to 2.04. This range of values is partially the effect of having skewed distributions but primarily result from outliers. To once again examine the effect on composite scores, additional averages were calculated. In Table 1 the column labeled "Average: This Min Z, Others Max" shows the average when fourteen scores were at the maximum possible and the other one at the minimum and are all converted to Z's. Observe that the highest composite value is 0.82 and the lowest composite is -0.10. This lowest composite score of -0.10 illustrates the potential strong effect of an outlier. For this case fourteen of fifteen scores were at the maximum, yet this composite score is below the composite score for all fifteen means. For a person having a raw score of 80 on variable 15 (possible range was 0 to 100) it was impossible to recover enough to reach the mean.

Obviously, even after converting these variables to the same mean and standard deviation, i.e. Z scores, there are considerable differences in the composite scores. These differences result from the effect of differences in the lowest value in each distribution, several of which are outliers.

In the context of the Tennessee Career Ladder, the values in the last column indicate that a person receiving the maximum score on fourteen of fifteen variables could possibly be evaluated as achieving Level I, Level II, or Level III, depending upon which particular variable had the low score. Also note that the lowest composite score occurs on a variable in which the lowest observed score was 80% of the highest possible score, but the distribution was extremely skewed.

In Tennessee the method used to alleviate such a strong influence of an outlier (or of a very skewed distribution) was to convert raw scores to Z scores and then any Z value below -3.00 was arbitrarily changed to -3.00 (also, any value above 3.00 became 3.00) before averaging to obtain a composite score. This was hoped to minimize the effect of an outlier on a variable when computing the composite score.

Situation 2: Time Series

In this situation, tolerance intervals were used to screen out outliers prior to performing time series analysis for real estate investment. The actual goal of that research was to determine whether an investor in real estate could predict future returns based upon past information. To determine this, a time series analysis testing for a pattern which could be used as a trading rule was performed. Prior to this analysis it was necessary to obtain stable estimates of monthly means of real estate investment per square foot. Consequently, it was necessary to remove the influence of outliers from each monthly mean. The method of identifying outliers was an adaptation of that described by Hawkins (1980).

Covering a 180 month period, approximately 129,000 transactions were

available for analysis, with the unit of measurement being sales price per square foot. These transactions were classified into ten planning districts to control for neighborhood effects upon price. For each month, analysis of variance was performed for the purpose of obtaining monthly means for each planning district and the mean square within, i.e. the pooled within group variance. Using the square root of the pooled within group variance as that month's standard deviation, for each district mean a tolerance interval was found which was the mean plus and minus two standard deviations. Each observation was compared to the tolerance interval for that month in that district. (There were 1800 such intervals.) If the observation was within the interval, it was selected for the sample; otherwise, it was removed from the study. About 122,000 observations were found to be within their respective tolerance interval and thus were selected from the 129,000 observations, for a 94.6 percent selection rate.

For each district, for each month, a revised mean was calculated based only on the observations selected as being within the respective tolerance interval. These revised monthly means for each district formed the data for the time series analyses to determine if a pattern was present in the analysis of return on investment. The time series demonstrated that in seven of the ten planning zones there was evidence of some "pattern" which might be used as a trading rule to achieve excess returns. Rayburn (1986) provides a detailed report of the results of these time series analyses.

Situation 3: Univariate Distribution

The purpose of this example was to demonstrate the effect of a single outlier on a univariate distribution. The data used was artificially created to represent a normal distribution with a mean of 500 and a standard deviation of 100. Three data sets were created with respective sample sizes of 20, 50, 100, and all of the distributions reflected the desired mean and standard deviation out to two decimal

places.

To demonstrate the effect of an outlier upon the mean and standard deviation of a set of normally distributed numbers, an extreme value was substituted for the minimum value in each data set. Means and standard deviations were recalculated with the new minimum values. Table 2 shows the results of this substitution for selected minimum values.

The original mean for each data set was 500. The presence of an extreme value reduced the mean for each set. When zero was substituted for the minimum value in a sample size of 20, the resulting mean was 484.9. When -500 was entered as an extreme value, the mean decreased to 459.9. Although the means also decreased for a sample sizes of 50 and 100, the effect of the outlier was lessened as the sample size increased.

Researchers have previously focused their attention on the effect of outliers on the mean of a distribution. However, the presence of an extreme value in the sample also had a marked effect on the standard deviation. As the mean for each sample was reduced, the standard deviation increased. When the sample size was 20, the inclusion of an extreme value of -500 reduced the mean by approximately 40 units. Yet, the increase in the standard deviation was almost 143%. As the sample size increased, the effect on the standard deviation was lessened but was still considerable. For all three illustrations, Hartley's F-max test of homogeniety of variance was significant at the .05 level. The effect of an outlier on a normal distribution is much more pronounced for the standard deviation than for the mean. Utilization of a mean and standard deviation from a distribution containing an outlier when performing a hypothesis test places a researcher in double jeopardy.

Situation 4: Maximum Correlation

This example demonstrated the effect of an outlier on the relationship of two

variables when one of the variates contains an extreme value. The same three normally distributed data sets from above were used to calculate correlation coefficients. Each data set was originally correlated with itself (r = 1.00). Then the correlation was determined between the original data set which was normally distributed and the same data set with an extreme value substituted for the minimum value. The various minimum values may be considered to be outliers in their respective univariate distributions. Correlation coefficients were then recalculated for each substitution on each of the different samples. Table 3 shows the results for this data analysis.

When the sample size was 20 and minimum value was replaced by 100, the correlation was reduced from 1.00 to .949. When the minimum value was -500, the correlation was .755. In the sample with 50 cases, the correlation was reduced from 1.00 to .801, and for a sample size of 100, it was reduced to .857. The presence of a single outlier on only one of the variates can have a sizeable effect on the relationship between the two variables. The effect of the outlier is less when the samples to be correlated contain a reasonably large number of cases. Note that for all of these distributions, if rank order correlations had been calculated all values would be 1.000.

Situation 5 : Selected Correlations

The purpose of this example was to illustrate the effect of an extreme value on two different bivariate distributions. In this situation, one variate was represented by the original normally distributed data. The normally distributed data for the other variate was rearranged twice so that the original correlation between the two variables was .707 (.500 shared variance) in one example and .316 (.100 shared variance) in the other with the additional constraint that the two minimum values were paired in each example. The minimum value on the originally ordered data was replaced by the various extreme values used above, and

correlations were recalculated. These various minimum values may be considered to be outliers in a univariate distribution. Results are presented in Tables 4 and 5.

The effect of the extreme value was different on each of the bivariate distributions. In the sample of twenty cases, when an extreme value was introduced into the situation where the correlation was .707, the correlation coefficient increased slightly (r = .723) and then gradually decreased (r = .634). This pattern also held true when the sample size was 100; however, the reduction in r was not as great. With a sample size of fifty the same pattern of increase followed by decrease is shown, but the decrease was more than when the sample size was twenty. This larger increase apparently results from the particular pattern of paired scores among the other values in the distribution.

The outliers had an extremely different effect in the example where the original correlation was .316. The presence of an extreme value on one of the variates markedly increased the correlation coefficient when the sample size was 20. By replacing the minimum value with a variety of extreme values, the correlation coefficient was increased to .474. With sample sizes of 50 and 100, there was also an increase in the correlation; however, it was not as dramatic.

Situation 6: Bivariate Outlier

The purpose of this example was to illustrate the effect of a bivariate outlier upon a correlation coefficient. In this situation, both variables were normally distributed with the data for one variable being rearranged so that the original correlation between the two variables was .707 in one example and .316 in the other, with the additional constraint in both examples that the two minimum scores were paired. When sample size was 20, the minimum value for the variable that was arranged in rank order was moved to the position of the fifth lowest value with those below being moved down one position and then the correlation

coefficient was calculated. Then the minimum value was moved the the tenth position, to the fifteenth position, and finally to the twentieth position with a correlation being calculated each time. Each time the minimum value was in a different position; however, the scores were still in the original rank order with the exception of that one value which was intentionally inserted out of sequence. When the sample size was 50 a similar procedure was used, but this time the minimum value was placed in the positions of the tenth, twentieth, thirtieth, fortieth, and fiftieth scores. For sample size of 50 there were six correlation coefficients calculated. When the sample size was 100, the minimum value was again substituted for every tenth value, so that eleven correlation coefficients were calculated.

Moving the minimum value through the range of values demonstrates the effect of varying the degree of a bivariate outlier. A bivariate outlier indicates a pair of scores with large deviation from the regression line. Table 6 shows these correlations when the original correlation was .707 and Table 7 shows them when the original correlation was .316. When the sample size was twenty, the change in value of the correlation was a dramatic decrease. When the initial correlation was .707, it decreased to .274; or in terms of shared variation the decrease was from .500 to .075. When the initial correlation was .316, it decreased to -.111, that is the direction of the relationship even changed. When the sample size was 50, there was also a considerable decrease in the correlation. When the initial correlation was .707, it decreased to .485; or in terms of shared variance the decrease was from .500 to .235. When the initial correlations was .316, it decrease to .095 (.100 to .009 shared variance). When the sample size was 100 the correlation of .707 decreased to .570 (.500 to .325 shared variance) and the correlation of .316 decreased to .180 (.100 to .032 shared variance). Thus, it can be concluded that one very extreme outlier has a considerable effect upon the calculation of a correlation even when both variables have univariate normal

distributions.

Summary

In summary, an outlier is an extreme value which has the potential to distort statistical results. The distortion effect of outliers when computing composite scores was demonstrated. The effect of an outlier in a distribution was shown to affect the mean, standard deviation, and correlation with other variables. The effect of a bivariate outlier, i.e. extreme deviation from the regression line, was also shown to reduce the estimate of the relationship between two variables.

TABLE 1. Statistics for Calculating Composite Score from Fifteen Variables after Standardizing

	Actual Range Min	Max	Mean	SD	Min Z	Average: This Min Z Others Mean	Max Z	Average: This Min Z Others Max
Var.1	0	1	0.50	0.50	-1.00	-0.07	1.00	.82
Var.2	0	1	0.70	0.46	-1.53	-0.10	0.65	.81
Var.3	0	1	0.90	0.30	-3.00	-0.20	0.33	.73
Var.4	0	1	0.99	0.10	-9.95	-0.66	0.10	.29
Var.5	1	5	3.00	1.27	-1.57	-0.10	1.57	.75
Var.6	1	5	4.00	0.93	-3.23	-0.22	1.08	.67
Var.7	1	5	4.50	0.66	-5.30	-0.35	0.76	.55
Var.8	1	5	4.80	0.60	-6.33	-0.42	0.33	.51
Var.9	1	5	4.90	0.54	-7.22	-0.48	0.19	.46
Var.10	1	9	5.00	1.96	-2.07	-0.14	2.04	.68
Var.11	1	9	8.00	1.80	-3.89	-0.26	0.56	.66
Var.12	11	20	17.00	1.80	-3.33	-0.22	1.67	.62
Var.13	1	100	50.00	25.00	-2.00	-0.13	2.00	.69
Var.14	1	100	80.00	10.00	-7.90	-0.53	2.00	.30
Var.15	80	100	99.92	1.26	-15.75	-1.05	0.06	-.10

TABLE 2. Means and standard deviations for univariate distributions.

	MEANS			STANDARD DEVIATIONS		
	N = 20	N = 50	N = 100	N = 20	N = 50	N = 100
Normalized (303,267,242)	500.0	500.0	500.0	100.0	100.0	100.0
Low 200	494.9	498.7	499.6	112.5	103.6	101.2
100	489.9	496.7	498.6	127.6	110.2	104.6
0	484.9	494.7	497.6	144.5	118.2	108.8
-100	479.9	492.7	496.6	162.7	127.2	113.8
-200	474.9	490.7	495.6	181.9	137.1	119.4
-300	469.9	488.7	494.6	201.7	147.7	125.5
-400	464.9	486.7	493.6	222.0	158.9	132.2
-500	459.9	484.7	492.6	242.7	170.5	139.2

TABLE 3. Maximum Correlations Between a Normal Distribution and Distributions Containing an Outlier

	Normalized		
Minimum Value	N = 20 (Min = 303)	N = 50 (Min = 267)	N = 100 (Min = 242)
Normalized	1.000	1.000	1.000
Low 200	.983	.996	.999
100	.949	.979	.991
0	.910	.954	.977
-100	.871	.923	.957
-200	.837	.891	.934
-300	.806	.860	.909
-400	.779	.829	.883
-500	.755	.801	.857

TABLE 4. Correlations Between a Normal Distribution and Distributions Containing an Outlier When Original Correlation is 0.707.

	Normalized		
Minimum Value	N = 20	N = 50	N = 100
Normalized	.707	.707	.707
Low 200	.723	.714	.710
100	.719	.714	.711
0	.707	.706	.708
-100	.691	.693	.700
-200	.675	.678	.689
-300	.660	.661	.676
-400	.647	.645	.662
-500	.634	.629	.647

TABLE 5. Correlations between a Normal Distribution and Distributions Containing an Outlier When Original Correlation is .316

	Normalized		
Minimum Value	N = 20	N = 50	N = 100
Normalized	.316	.316	.316
Low 200	.376	.336	.323
100	.413	.359	.338
0	.437	.375	.349
-100	.452	.386	.356
-200	.461	.393	.361
-300	.467	.397	.364
-400	.471	.399	.366
-500	.474	.400	.366

TABLE 6. Correlation Coefficients for Two Normal Distributions in Presence of Bivariate Outliers

Position of Minimum Value	N = 20	N = 50	N = 100
1	.707	.707	.707
5	.620		
10	.560	.669	.688
15	.462		
20	.274	.644	.679
30		.622	.673
40		.586	.667
50		.485	.660
60			.653
70			.646
80			.635
90			.621
100			.570

TABLE 7. Correlation Coefficients for Two Normal Distributions in Presence of Bivariate Outliers

Position of Minimum Value	N = 20	N = 50	N = 100
1	.316	.316	.316
5	.199		
10	.119	.260	.290
15	.069		
20	-.111	.230	.278
30		.202	.270
40		.178	.263
50		.095	.256
60			.250
70			.243
80			.233
90			.221
100			.180

REFERENCES

Bancroft, T. A., & Han, C. (1981). Statistical Theory and Inference in Research. New York: Marcel Kekker.

Barnett, V. & Lewis, T. (1978). Outliers in Statistical Data. New York: John Wiley & Sons.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. Annals of Mathematical Statistics, 21, 27-58.

Hawkins, D. M. (1980). Identification of Outliers. New York: Chapman and Hall.

Killeen, P. R. (1985). The bimean: A measure of central tendency that accommodates outliers. Behavior Research Methods, Instruments, & Computers, 17, 526-528.

Lyon, A. J. (1970). Dealing with Data. New York: Pergamon Press.

Naus, J. I. (1975). Data Quality Control and Editing. New York: Marcel Dekker.

Rayburn, W. B. (1986). Application of the Efficient Market Hypothesis to Residential Real Estate Returns. Unpublished doctoral dissertation, Memphis State University, Memphis.

Rey, W. J. J. (1983). Introduction to Robust and Quasi-Robust Statistical Methods. New York: Springer-Verlag.

Sachs, L. (1984). Applied Statistics: A Handbook of Techniques (2nd.ed.). New York: Springer-Verlag.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Winer, B. J. (1971). Statistical Principles in Experimental Design (2nd.ed.). New York: McGraw-Hill Book Co.